PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Cohort Profile: A Data Linkage Cohort to Examine Health Service Profiles of People with Intellectual Disability in New South Wales, Australia
AUTHORS	Reppermund, Simone; Srasuebkul, Preeyaporn; Heintze, Theresa; Reeve, Rebecca; Dean, Kimberlie; Emerson, Eric; Coyne, David; Snoyman, Phillip; Baldry, Eileen; Dowse, Leanne; Szanto, Tracey; Sara, Grant; Florio, Tony; Trollor, Julian

VERSION 1 – REVIEW

REVIEWER	Emeritus Professor Roy McConkey Institute of Nursing and Health Research, Ulster University. N. Ireland, United Kingdom
REVIEW RETURNED	26-Oct-2016

GENERAL COMMENTS	The authors rightly present the arguments for Big Data studies that link the information about individuals which is held in various administrative datasets. I applaud their brave attempt to do this for a population of people labelled as having intellectual disabilities. Not only is this a small minority of health service users. it is also a very heterogeneous. To my mind, the main value of this study is not so much in describing the resultant cohort but rather providing insights into the process of creating the cohort, unpacking the decision making and evaluating the product against the effort entailed in its production. The authors seem convinced of the value of the cohort they have identified but I am less persuaded. The title of the paper is potentially misleading. Health status and service usage will be defined solely in terms of data obtained from acute hospitals. There is no mention at all of community health services and the yet to be realised potential of linking data held on GP records. Of course this raises the issue as to where efforts in data linking are deployed. Do we advocate for access to arguably more relevant datasets or do we make do with whatever is available? of course it need not be a choice but the authors surely have a view based on an assessment of the extent to which they have met their aims. The authors gave no insight into the time and effort they expended in creating the cohort. What were the obstacles and how were they overcome? Perhaps some were not resolved which would suggest certain cautions around how the data is interpreted or used. At various points the authors assert ICD 10 criteria were used to identify ID cases. I could accept this might happen for the disability

	dataset although my experiences in analysing national datasets suggests that this more a hope than a reality. However I am much less convinced that personnel in a busy ED will take cognisance of ICD 10. Likewise the number of different codes in the acute hospital dataset (over 60) suggests that other criteria influence the code assigned to a person. This uncertainty has particular relevance to persons identified only on the two hospital datasets. The authors rightly acknowledge the potential of this approach in identifying persons with ID who do not receive specialist services. This raises the issue as to how improvements may be needed to recording information on existing datasets. The authors seem to be content to take what is there, albeit with imperfections. I struggle to understand how an administratively linked dataset will lead to improved models of education as per the second bullet point under strengths on page 4. It might further provide a rationale for so doing but other research and development projects will be required to achieve this outcome. indeed the listings of influences on the health and well-being of people with ID which the authors presented in the introduction would support this. Once data linkage was achieved it seemed that only 34 percent of the identified persons with ID would have linked data across the three datasets. As figure 1 shows, other persons were linked across two or appeared on only one dataset. What are the implications then for data analysis and is 34 percent a good outcome for a data linking project? Moreover there appear to be no plans to update the cohort over time so as to track health interventions and possible outcomes. Is it possible to create a 'live' linked database? The second cohort was persons with mental health admissions who formed an equally heterogeneous population with 80 codes identifying them for this study. I am puzzled by the different totals presented for people with ID in figure 2 and figure 1. I note that for this population only 2 percent were matched on all three datasets. The authors end the article with a listing of possible analyses they intend to undertake. However I hope this would not deflect from further research and improvements to the data linking undertaken thus far and broadening the range of health data that is needed to ultimately achieve better health outcomes for persons with intellectual disability.

REVIEWER	Rory Sheehan University College London, UK
REVIEW RETURNED	27-Oct-2016

GENERAL COMMENTS	Cohort profile: a data linkage cohort to examine the health status and service use of people with intellectual disability in New South Wales, Australia GENERAL This is important work which has the potential to improve understanding of the health needs of people with ID and influence services. I'm sure much good will come of this. My main criticism in the way it is written is that the authors consider the work as very

	much part of a larger programme of research – this is very well but the paper should stand alone. The authors also make quite strong and sweeping statements about the impact of the work which I do not necessarily feel are supported by the data that are presented here. I was surprised to see not a discussion but a "future directions" section at the end of the paper. The plans seem laudable, but I did not think this was a protocol paper. There is obviously lots of good work in the pipeline but I'm not sure whether where this paper sits. Maybe the authors want a paper that will outline the methodology and basic socio-demographic data which they can later cite? If so, this paper is only half that. Alternatively they may wish to present a protocol of their work – but again, this is not it. The paper is well over the suggested word count of 4,000. ABSTRACT The abstract is not in house style. I think the focus of the work could be explained better by some minor changes to the wording. The purpose is to "understand the interaction of health and disability services" – in the next sentence we hear that it will inform the development of improved "health and mental health services" and in the next section "a focus of this work is on mental health". INTRODUCTION Well written and sets the case for the study clearly. Page 6, line 55 to page 7 line 5 – this sentence does not quite make sense to me "…and the service pathways are engaged". Page 6, line 11 – can the authors reference the "substantial unmet health needs"? COHORT DESCRIPTION (should this not be "METHODS"?) Page 8, line 23 – "all people identified as having an intellectual disability meet the specific requirements for a DSM IV or ICD-10 diagnosis" – this is quite a bold statement – could the authors clarify or temper? See also page 8, line 54 to page 9, line 6 – how was this measured? Page 8, line 49 – did the cohort of people with ID cover all those with ID in the database, or is this a sample? If a sample, how was this chosen to be representative of the whole? It is useful to have the variables included in each database as supplementary data. 5% of the sample is defined as having an ID – would the authors care to comment on this, as the figure quoted in population prevalence studies is usually much less than this. I note that people with autism without ID have not been included, so where it the 'excess' coming from? Or is this an enriched population? I do wonder whether the comparison cohort are the right comparison

	– people with long-term mental health problems also experience higher rates of comorbid illness and might be subject to many of the same barriers to healthcare / discrimination as those with ID so to use this population as a benchmark for health seems up for debate. Most of the references are to papers which have compared people with ID against non-ID counterparts, regardless of mental health status. FINDINGS TO DATE ("RESULTS") As a person may have multiple records in the full analysis, can the authors confirm that they were uniquely identified and not counted multiple times? The databases are obviously rich in detail and the authors present many interesting results. I wonder though if this part of the manuscript could be re-written for readability and accessibility? I think we also see a slight lack of focus in the aims of the work presented in this paper as some of the data that are highlighted in the text seem rather random – I would support the authors making fewer, but stronger points which can be interpreted and discussed at more length. For example, there is emphasis on the number living in cities – differences seem minor and the relevance of this is not interpreted later. FUTURE DIRECTIONS I can't really comment on this further – it is a plan of a work programme. The strengths and limitations seem reasonably comprehensive, but again refer to work which has not yet been done! I would like to see a concise paragraph at the end which very neatly summarises the paper and its clinical implications.

VERSION 1– AUTHOR RESPONSE

Reviewer 1:

1. The authors rightly present the arguments for Big Data studies that link the information about individuals which is held in various administrative datasets. I applaud their brave attempt to do this for a population of people labelled as having intellectual disabilities. Not only is this a small minority of health service users. it is also a very heterogeneous. To my mind, the main value of this study is not so much in describing the resultant cohort but rather providing insights into the process of creating the cohort, unpacking the decision making and evaluating the product against the effort entailed in its production. The authors seem convinced of the value of the cohort they have identified but I am less persuaded.

   We have now added more details of how the cohort was created, its challenges, value and shortcomings (see also reply to comment 3) and included a new paragraph "Project Resourcing and Development" (page 7) to provide more insights into the resources required to run this study. We are convinced of the value of our data linkage, with all the mentioned shortcomings, as interrogation of these data will give insights into the diagnostic and service use profile of people with ID. Furthermore, there is limited Australian data that examines the prevalence and impact of mental disorders, the impact on support persons, and the direct cost to health services. The analysis of linked health and disability service data will allow us to develop a sound epidemiological and service evidence base that will inform our understanding of service level usage; pathways through the service system taken by people with an ID; and barriers and

enablers of access to care. It will achieve this by interrogating the linked datasets and triangulating this with the data derived from an analysis of Commonwealth and State Mental Health Policy and a qualitative research approach with stakeholder engagement to improve accessibility. We have added this to the "strength and limitations" paragraph.

2. The title of the paper is potentially misleading. Health status and service usage will be defined solely in terms of data obtained from acute hospitals. There is no mention at all of community health services and the yet to be realised potential of linking data held on GP records. Of course this raises the issue as to where efforts in data linking are deployed. Do we advocate for access to arguably more relevant datasets or do we make do with whatever is available? of course it need not be a choice but the authors surely have a view based on an assessment of the extent to which they have met their aims.

   We did not include community health services information, including GP visits, in this study as it requires linking data from a different jurisdiction and this was not feasible at the time of this study. Further, the health services information generally contains no clinical information as administrative datasets are mainly used for payment purposes. Nevertheless, we believe that the available data linkage is of great value. The use of administrative data in research projects allows the development of appropriate resources and policy which ameliorates the impact of diseases in the community. Our study aims to develop an epidemiological profile related to the health and wellbeing of people with ID. Currently direct service system information about the health and mental health care needs of people with ID is inaccessible, and locked within administrative data sets of relevant government agencies (disability, health, education, corrections, etc). Our project aims to improve the knowledge base by interrogating linked service system data related to this population group. The project has a strong translational component, and results will be used to inform policy and services development in this area. We have changed the title to "A Data Linkage Cohort to Examine Health Service Profiles of People with Intellectual Disability in New South Wales, Australia" and we have added the lack of community health services/GP data as a limitation (page 20).

3. The authors gave no insight into the time and effort they expended in creating the cohort. What were the obstacles and how were they overcome? Perhaps some were not resolved which would suggest certain cautions around how the data is interpreted or used.

   We have added information about the time and effort expended in creating the cohort and the obstacles as well as how we dealt with them (page 7 and page19).
   In particular a study like this is very resource intensive (financially, personnel and time consuming to apply for and to combine and clean the datasets). We were reliant on data custodians from different organisations to release the data in a timely manner and the process of obtaining ethics approval and receiving the linked data was slower than anticipated.

4. At various points the authors assert ICD 10 criteria were used to identify ID cases. I could accept this might happen for the disability dataset although my experiences in analysing national datasets suggests that this more a hope than a reality. However I am much less convinced that personnel in a busy ED will take cognisance of ICD 10.
   Likewise the number of different codes in the acute hospital dataset (over 60) suggests that other criteria influence the code assigned to a person. This uncertainty has particular relevance to persons identified only on the two hospital datasets. The authors rightly acknowledge the potential of this approach in identifying persons with ID who do not receive specialist services. This raises the issue as to how improvements may be needed to recording information on existing datasets. The authors seem to be content to take what is there, albeit with imperfections.

5. *The reviewer raises an important point regarding the accuracy of diagnostic coding within administrative datasets. Coding of all diagnoses, including ID, is routinely completed for NSW health services. However, we are unable to verify the completeness or accuracy of such coding. ID identification from ED as it is less likely to identify people from ED data if ID was not the main reason for their visit. This point is evident in Figure 1. This is mentioned in the limitations paragraph (page 19), and as mentioned by the reviewer, improvements are needed to the recording of such information. Fortunately, as recognised by the reviewer, the substantial majority of our ID cohort is identified by the DS-MDS, which uses ID diagnosis as a requirement*

for service entry. This, plus the robust proportion of the NSW state population captured as having ID, suggests that a sizeable proportion of the ID population is captured in our cohort. We intend to examine the level of agreement between ID diagnostic coding within the different service system compartments, and make recommendations regarding improvements. This will be the topic of an additional paper.

6. I struggle to understand how an administratively linked dataset will lead to improved models of education as per the second bullet point under strengths on page 4. It might further provide a rationale for so doing but other research and development projects will be required to achieve this outcome. indeed the listings of influences on the health and well-being of people with ID which the authors presented in the introduction would support this.

   We agree that the interrogation of administrative datasets alone does not lead to improved training and education and we have removed this statement. However, as mentioned on page 7, the data linkage work is part of an interdisciplinary and collaborative project including policy analysis, a qualitative research approach to identify barriers and enablers to accessing mental health services as well as a comprehensive knowledge translation framework to translate the findings into policy and practice. Therefore, our data linkage findings can and will be used to inform the development of appropriate training and education for services providers.

7. Once data linkage was achieved it seemed that only 34 percent of the identified persons with ID would have linked data across the three datasets. As figure 1 shows, other persons were linked across two or appeared on only one dataset. What are the implications then for data analysis and is 34 percent a good outcome for a data linking project? Moreover there appear to be no plans to update the cohort over time so as to track health interventions and possible outcomes. Is it possible to create a 'live' linked database?

   Figure 1 shows the number of people with ID from multiple service datasets. It is intended to be read that the majority of people with ID were identified from DS-MDS, followed by the APDC and the EDDC datasets. Overall we have 51,452 people with ID, 42,243 from the DS-MDS. The overlap across the three datasets here is irrelevant. To be included in our cohort, each individual had to receive a service with an ID flag and not all people with ID would also have a hospital admission or ED presentation. This has now been clarified in the cohort definition paragraph on pages 10-11.

   We are currently in a process of updating our cohort; it will include more datasets and a longer follow up period (please see also response to comment 8). This is now mentioned on page 20. If by the 'live' linked database you mean real time linkage, it is impossible to do so in the Australian system. As aforementioned, administrative data collections in Australia are for payment purposes and thus don't reflect real time updates.

8. The second cohort was persons with mental health admissions who formed an equally heterogeneous population with 80 codes identifying them for this study. I am puzzled by the different totals presented for people with ID in figure 2 and figure 1. I note that for this population only 2 percent were matched on all three datasets.

   The 2 percent matched on three datasets reflects the fact that people with mental health admissions by itself are not qualified to receive disability services, unless they have an additional disability. As per reviewer 2 comment #12, we have decided to not emphasise the second cohort in this paper.

9. The authors end the article with a listing of possible analyses they intend to undertake. However I hope this would not deflect from further research and improvements to the data linking undertaken thus far and broadening the range of health data that is needed to ultimately achieve better health outcomes for persons with intellectual disability.

   The "future directions" paragraph has been considerably shortened. In the strength and limitations section we now point out how we will broaden the range of the data. Specifically, we will add data from Corrective Services NSW, NSW Department of Education and NSW Public Guardian and we will extend the timeframe to 2001-2016. This will allow us to identify, quantify

and cost health and other services provision to people with ID within the various cohorts of interest. This is now mentioned on page 20.

Reviewer 2:

GENERAL

1. This is important work which has the potential to improve understanding of the health needs of people with ID and influence services. I'm sure much good will come of this. My main criticism in the way it is written is that the authors consider the work as very much part of a larger programme of research – this is very well but the paper should stand alone. The authors also make quite strong and sweeping statements about the impact of the work which I do not necessarily feel are supported by the data that are presented here.

   The manuscript is written as a "cohort profile". The rationale for this type of paper is summarised on the BMJ Open website (instruction for authors) as follows: "The cohort profile is an article type set up in BMJ Open to fill the space between a study protocol and a results paper. Cohort profiles should describe the rationale for a cohort's creation, its methods, baseline data and its future plans. Cohorts described should be long-term, prospective projects and not time-limited cohorts established to answer a small number of specific research questions."

2. I was surprised to see not a discussion but a "future directions" section at the end of the paper. The plans seem laudable, but I did not think this was a protocol paper. There is obviously lots of good work in the pipeline but I'm not sure whether where this paper sits. Maybe the authors want a paper that will outline the methodology and basic socio-demographic data which they can later cite? If so, this paper is only half that. Alternatively they may wish to present a protocol of their work – but again, this is not it.

   As mentioned in the previous response, our manuscript is not a protocol paper but a cohort profile. The reviewer is right that we want a cohort profile paper that outlines the methodology and basic socio-economic data to cite as a basis for future outputs. However, we made some major changes to the paper (e.g. shortened the 'future directions' paragraph and shortened the 'findings to date' section) in order to make it more succinct and more recognisable as a cohort profile paper.

3. The paper is well over the suggested word count of 4,000.

   The author guidelines for cohort profiles do not mention a word limit. However, we have shortened the manuscript considerably and it now contains 4274 words.

ABSTRACT

4. The abstract is not in house style.

   We followed the author guidelines for cohort profile articles: "Use these headings to provide brief descriptions of the following:
   Purpose: describe why the cohort was set up
   Participants: describe who is in the cohort
   Findings to date: what data has been collected so far and any major results
   Future plans: how will the cohort be used in future, including any date for completion of data collection"

5. I think the focus of the work could be explained better by some minor changes to the wording. The purpose is to "understand the interaction of health and disability services" – in the next sentence we hear that it will inform the development of improved "health and mental health services" and in the next section "a focus of this work is on mental health".

   We thank the reviewer for pointing this out. The focus of the work is to interrogate a large linked dataset to provide evidence which will inform the development of improved health and mental health services for people with ID. The results of this study will be used to inform the development of health and mental health services for people with ID. A specific subtheme of this

research is the development of a detailed understanding of their representation in the mental health components of the data. This has now made clearer throughout the paper.

INTRODUCTION

6. Well written and sets the case for the study clearly.
   Page 6, line 55 to page 7 line 5 – this sentence does not quite make sense to me "…and the service pathways are engaged".

   We reworded this sentence to "…and the service pathways that have been used."

7. Page 6, line 11 – can the authors reference the "substantial unmet health needs"?

   We have included Evans et al. Journal of intellectual disability research, 2012. 56(11): p. 1098-1109 and van Schrojenstein Lantman-De Valk et al. Fam Pract, 2000. 17(5): p. 405-7 as references.

8. COHORT DESCRIPTION (should this not be "METHODS"?)

   The author guidelines for cohort profiles require a "cohort description" section rather than a methods section.

9. Page 8, line 23 – "all people identified as having an intellectual disability meet the specific requirements for a DSM IV or ICD-10 diagnosis" – this is quite a bold statement – could the authors clarify or temper? See also page 8, line 54 to page 9, line 6 – how was this measured?

   Fulfilment of DSM IV criteria for intellectual disability was required in order to be eligible to receive a service due to intellectual disability. For those who did not receive disability services from ADHC, we used ICD-10 as recorded in APDC and EDDC.

10. Page 8, line 49 – did the cohort of people with ID cover all those with ID in the database, or is this a sample? If a sample, how was this chosen to be representative of the whole?

    Our cohort covered all people with ID who received disability services or health services. Our cohort is therefore representative to people with ID who received disability services, health services or both. The cohort is representative for people with moderate to severe ID, however, our ID cohort accounts for 0.6% of the NSW population in 2011 and people with mild ID may be underrepresented. This has been added to the 'cohort definition' paragraph.

It is useful to have the variables included in each database as supplementary data.

11. 5% of the sample is defined as having an ID – would the authors care to comment on this, as the figure quoted in population prevalence studies is usually much less than this. I note that people with autism without ID have not been included, so where it the 'excess' coming from? Or is this an enriched population?

    The 5% of the sample identified as having an ID refers to people with mental ill health who also had an ID. This is in accordance with previous studies showing that people with ID have higher rates of mental health issues (e.g. Cooper et al. (2007). Mental ill-health in adults with intellectual disabilities: prevalence and associated factors. The British Journal of Psychiatry,190(1): 27-35.). As mentioned above, our ID cohort accounted for 0.6% of NSW population in 2011.

12. I do wonder whether the comparison cohort are the right comparison – people with long-term mental health problems also experience higher rates of comorbid illness and might be subject to many of the same barriers to healthcare / discrimination as those with ID so to use this population as a benchmark for health seems up for debate. Most of the references are to papers which have compared people with ID against non-ID counterparts, regardless of mental health status.

    The reviewer rightly points out this issue. The cohort of people with mental ill health will only be used in the mental health related studies and not for general health studies. In addition, people

with ID, people with mental health issues and people with ID and mental health issues will be compared with published data from the general population in NSW. We have added this information to the cohort definition paragraph.

FINDINGS TO DATE ("RESULTS")

13. As a person may have multiple records in the full analysis, can the authors confirm that they were uniquely identified and not counted multiple times?

    After the linkage process, the CHeReL set up a unique person number for each individual which can be used to identify the same person across multiple datasets. This is explained in the 'Data Linkage' paragraph.

14. The databases are obviously rich in detail and the authors present many interesting results. I wonder though if this part of the manuscript could be re-written for readability and accessibility? I think we also see a slight lack of focus in the aims of the work presented in this paper as some of the data that are highlighted in the text seem rather random – I would support the authors making fewer, but stronger points which can be interpreted and discussed at more length. For example, there is emphasis on the number living in cities – differences seem minor and the relevance of this is not interpreted later.

    We have revised the "findings to date" section and shortened the "future directions" section to increase readability.

FUTURE DIRECTIONS

15. I can't really comment on this further – it is a plan of a work programme.

    We have now shortened the future directions paragraph and focused more on the findings to date.

16. The strengths and limitations seem reasonably comprehensive, but again refer to work which has not yet been done!

    As this is a cohort profile paper rather than a research paper, we refer to work which will be done with these data. However, we also focus on the linked dataset/cohort, e.g. the limited coverage, false-positive and false-negative links and heterogeneity.

17. I would like to see a concise paragraph at the end which very neatly summarises the paper and its clinical implications.

    We have added a concluding paragraph at the end (pages 20-21).

VERSION 2 - REVIEW

REVIEWER	Rory Sheehan University College London, England
REVIEW RETURNED	25-Jan-2017

GENERAL COMMENTS	This paper makes much more sense to me as a Cohort Profile, with substantial reduction in length, and with the amendments made as suggested by previous reviews.

REVIEWER	Professor Roy McConkey Institute of Nursing and Health Research, Ulster University, Northern Ireland.
REVIEW RETURNED	27-Jan-2017

GENERAL COMMENTS	I welcome the revisions that the authors have made to this paper and the clarification that it is a submission as a cohort study as defined by the Journal. There are a number of issues that the authors have not addressed or still need to be addressed. It seems there is little prospect that linkages with primary health care records will be attained in the foreseeable future. Hence the title - but certainly the abstract and the section on strengths and limitations - needs to specify that the health service is essentially hospital service usage. Moreover the lack of information on primary health care usage needs to be specified as a limitation in the summary on page 4. The number of cases identified in the Disability Services data (DS-MDS) should be added to the description. (On data given later I calculate this to be 73,674). Within the text it should be made clear this dataset covers children and adults. The authors intend to compare the cohort's other service utilisation to general population statistics (p.12) and standardise by age and sex. But given the link between ID and deprivation reported in the cohort, can the comparison be standardised also by deprivation? I am puzzled by the numbers presented in Figure 1. To what do the numbers in square brackets refer under the three main datasets as there is a discrepancy between this total and the numbers given the subsections for that database? For example: EDDC total is given as 24,242 but the three subgroups add to 22,344. In addition we are told that 29,902 records on the DS-MDS did not link to any other data base (p.12) yet Figure 1 states it is 18,142. What is the explanation for this discrepancy? Fully linked data is available for 17,267 cases which represents 23% of the persons availing of disability services in NSW over the chosen time period (n=73,674) which reduces to 21% if the additional people identified as ID on the EDDC and APDC are added (n=81,748). If the authors accept these figures then they should be included in the paper so that readers are aware that any findings emerging from future analyses are based on a small subset of the potential population. Table 1 does not present comparable data obtained from the DS-MDS which would provide readers with a useful comparator for comparisons with persons with ID using inpatient and emergency hospital services. On page 19 lines 31-33: the authors state that the data has been collected for administrative rather than clinical purposes and as such has significant shortcomings. It would be helpful for the authors to give examples of particular shortcomings. On page 20: lines 17-22; the author state that "some variables, for example, relating to severity of disability or measures of adaptive behaviour, that we would like to include in our models are not available in the data". Yet Table 1 in the Appendix indicates that this information is collected in the DS-MDS data.

	On page 21: line 28: the authors note that analysis of linked data is currently authorised to occur at only one location, owing to ethical considerations. Is this not a further limitation of this particular cohort? Perhaps the authors should expand on the ethical issues they have encountered in seeking to undertake the data linking (p. 13). An estimate of the total amount of monies expended in creating the cohort is not given although we can guess at it. I am left wondering the extent to which this approach will achieve the aims set for it and its overall cost-effectiveness but I appreciate that this will only become apparent in future years. Nevertheless the paper contains salutatory lessons for researchers in other countries who may wish to emulate this approach.

VERSION 2 – AUTHOR RESPONSE

We would like to thank Professor Roy McConkey for his helpful comments and suggestions, each of which have been addressed below.

1. It seems there is little prospect that linkages with primary health care records will be attained in the foreseeable future. Hence the title - but certainly the abstract and the section on strengths and limitations - needs to specify that the health service is essentially hospital service usage. Moreover the lack of information on primary health care usage needs to be specified as a limitation in the summary on page 4.

We have now specified in the abstract that our linked data contain hospital admission and emergency department presentations and added to the limitations summary on page 4 that it does not contain primary health care records. In the limitations section on page 20 it is written that the current linkage does not include community health services or GP records.

2. The number of cases identified in the Disability Services data (DS-MDS) should be added to the description. (On data given later I calculate this to be 73,674). Within the text it should be made clear this dataset covers children and adults.

We have added the total number of cases in the DS-MDS (n= 73,674) in the description of the database on page 8-9 and specified that it covers children and adults.

3. The authors intend to compare the cohort's other service utilisation to general population statistics (p.12) and standardise by age and sex. But given the link between ID and deprivation reported in the cohort, can the comparison be standardised also by deprivation?

This is a good point, however, it is more difficult to standardise by deprivation using the direct method as we do not have the details in the standard population. If published data permits, we will adjust for the impact of deprivation using the regression method. This has been added to the paragraph on page 12.

4. I am puzzled by the numbers presented in Figure 1. To what do the numbers in square brackets refer under the three main datasets as there is a discrepancy between this total and the numbers given the subsections for that database? For example: EDDC total is given as 24,242 but the three subgroups add to 22,344.

We have revised figure 1 to make it easier to read. The figure now shows the number of people who have a record exclusively in the APDC in green (n= 6,136), exclusively in the EDDC in blue (n= 40),

exclusively in the DS-MDS in red (n= 18,142), those who have a record in the DS-MDS as well as in the APDC in red/green (n= 2932), those who have a record in the DS-MDS as well as in the EDDC in red/blue (n= 5037), those who have a record in the APDC as well as in the EDDC in green/blue (n= 1898) and those who have a record in all 3 datasets in red/green/blue (n= 17,267). We have further added a footnote to the table explaining the formation of the cohort with ID.

To clarify, in your example for the EDDC, you added those who have a record in all 3 databases (n= 17,267) plus those who have a record in the EDDC as well as in the DS-MDS (n= 5,037) plus those who have a record exclusively in the EDDC (n= 40). However, you did not add those who have a record in the EDDC as well as in the APDC (n= 1,898).

5. In addition we are told that 29,902 records on the DS-MDS did not link to any other data base (p.12) yet Figure 1 states it is 18,142. What is the explanation for this discrepancy?

Figure 1 shows the numbers for people with ID only while the text on page 12 refers to all people in the DS-MDS dataset (i.e. with and without ID). This is clearer now in the revised figure 1.

6. Fully linked data is available for 17,267 cases which represents 23% of the persons availing of disability services in NSW over the chosen time period (n=73,674) which reduces to 21% if the additional people identified as ID on the EDDC and APDC are added (n=81,748). If the authors accept these figures then they should be included in the paper so that readers are aware that any findings emerging from future analyses are based on a small subset of the potential population.

The 17,267 people are those who have a record simultaneously in all 3 datasets. This is a different concept to 'linkage', and all individuals with ID have had linkage completed to the two health datasets (APDC, EDDC) and two mortality datasets. If individuals did not use emergency or inpatient facilities or did not die, they will not appear in those particular datasets. For example, the total population to examine the rates of hospitalisation in people with ID is 51,452 but there are only 28,233 individuals with at least one hospital admission. The overall rate will be calculated from total number of admissions from 28,233 individuals divided by person times of 51,452 individuals.

7. Table 1 does not present comparable data obtained from the DS-MDS which would provide readers with a useful comparator for comparisons with persons with ID using inpatient and emergency hospital services.

We have added a column containing DS-MDS data to Table 1.

8. On page 19 lines 31-33: the authors state that the data has been collected for administrative rather than clinical purposes and as such has significant shortcomings. It would be helpful for the authors to give examples of particular shortcomings.

Examples of particular shortcomings have been added to page 19. Significant shortcomings are that these datasets do not contain clinical information or information about the severity of the disability. For example in the hospitalisation data, if a person had multiple diagnoses in one episode, we do not have information on the length of each diagnosis or the severity of it. This has been added to page 19.

9. On page 20: lines 17-22; the author state that "some variables, for example, relating to severity of disability or measures of adaptive behaviour, that we would like to include in our models are not available in the data". Yet Table 1 in the Appendix indicates that this information is collected in the DS-MDS data.

The DS-MDS lists the primary and secondary disability group and different support needs. However, it does not contain information about the severity of ID (mild, moderate, severe or profound) or measures of adaptive behaviour.

10. On page 21: line 28: the authors note that analysis of linked data is currently authorised to occur at only one location, owing to ethical considerations. Is this not a further limitation of this particular cohort? Perhaps the authors should expand on the ethical issues they have encountered in seeking to undertake the data linking (p. 13).

The linked disability data is regarded as sensitive data and it is common practice that only one research unit is allowed to host such sensitive data. However, this does not prevent future collaborative work, as potential collaborators can work closely with our data analysts and/or in case of local collaborators, they could be based at our Department after approval from the ethics committee. We have added the ethical restriction to the Ethics paragraph on page 13: “An ethics requirement is that the linked data can only be analysed at the physical location of the researchers undertaking this work (The Department of Developmental Disability Neuropsychiatry, UNSW Sydney).”

11. An estimate of the total amount of monies expended in creating the cohort is not given although we can guess at it. I am left wondering the extent to which this approach will achieve the aims set for it and its overall cost-effectiveness but I appreciate that this will only become apparent in future years. Nevertheless the paper contains salutatory lessons for researchers in other countries who may wish to emulate this approach.

We agree that a research project like this is cost- and resource intensive, and have previously listed the resources required to develop and conduct the project under the heading “Project Resourcing and Development“ on page 7-8 of the manuscript. However, once the data is available and clean, it provides an excellent opportunity to examine research questions that could not be answered otherwise. An example of the power of this approach is a recent publication in BMJ Open http://bmjopen.bmj.com/content/7/2/e013489, which, one week after publication has already had an extensive impact on health advocacy for people with ID in Australia.

VERSION 3 – REVIEW

REVIEWER	Professor Roy McConkey Institute of Nursing and Health Research, Ulster University, Northern Ireland, United Kingdom
REVIEW RETURNED	06-Mar-2017

GENERAL COMMENTS	The clarifications provided by the authors are welcome and apologies for any misunderstanding on my part. I hope further analyses of the linked data lives up to the authors' expectations as this would certainly assuage my concerns about the cost-effectiveness of using Big Data with specialised populations in one geographical and the possible lack of suitable comparisons with the non-disabled population.